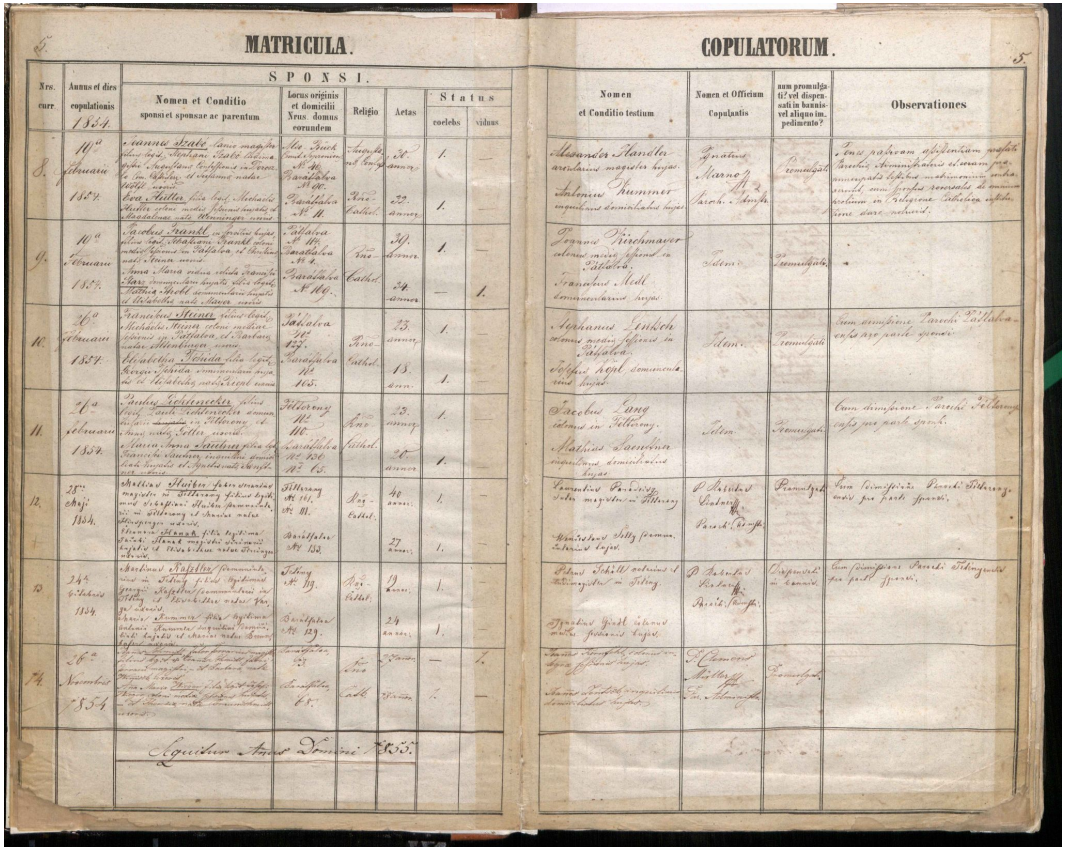
Digitising Handwritten Parish Records using Deep Learning

SoDa Labs - Monash Business School

How does the data look like?

Overview

- Birth, baptism, marriage, and death records from Austria (and Germany)
- Period: 1625-1938
- 183,000 church registers from Austria; ~4.5 million images



A church register from Austria, 1855 (source: https://data.matricula-online.eu/)

Aim

To digitise tabular data contained in raw images (unstructured data) in the form of spreadsheet (structured data)

2 sub-aims:

- Layout analysis
- Optical character recognition (OCR)

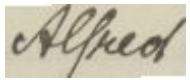
Aim: Layout Analysis

To discern structure of tables found in parish registers images:

- Identify field names
- Find separator lines (for segregating rows and columns)
- Crop snippets of individual text-entities to be used in OCR
- Position each text-entity into its corresponding row and column

Aim: Optical character recognition (OCR)

To parse handwritten text imprinted on the images to produce machine-readable text:



Alfred

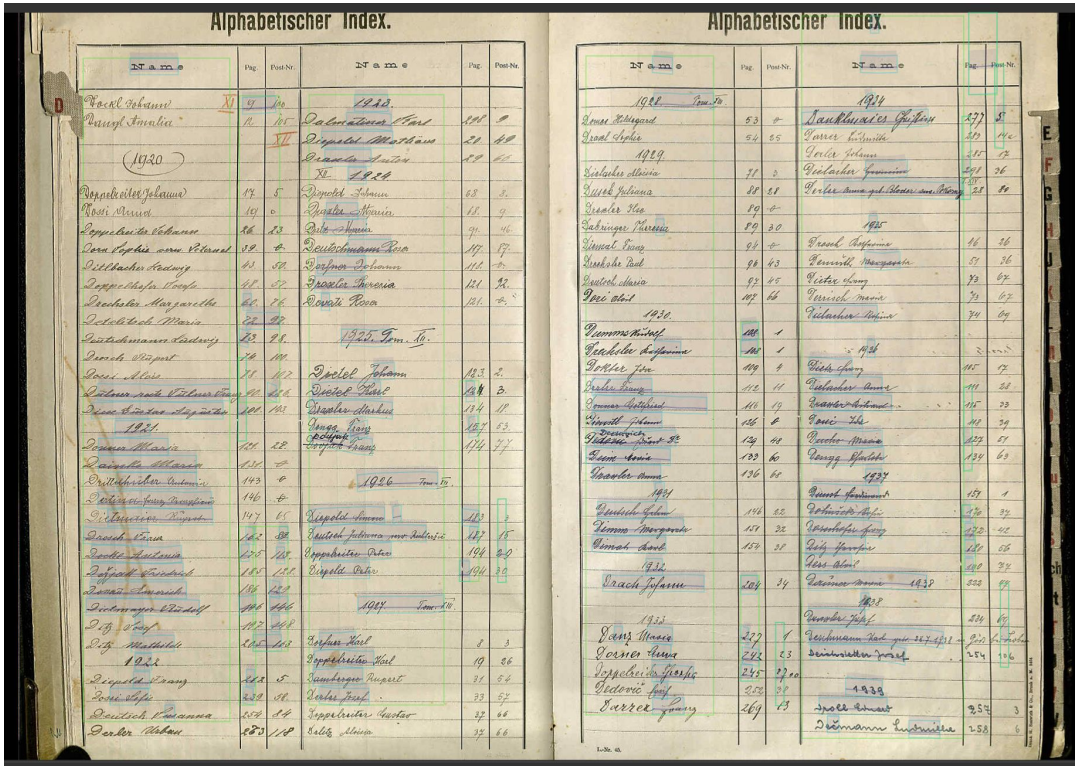
Maria

Lorenz

1918

Off-the-shelf solutions don’t always work

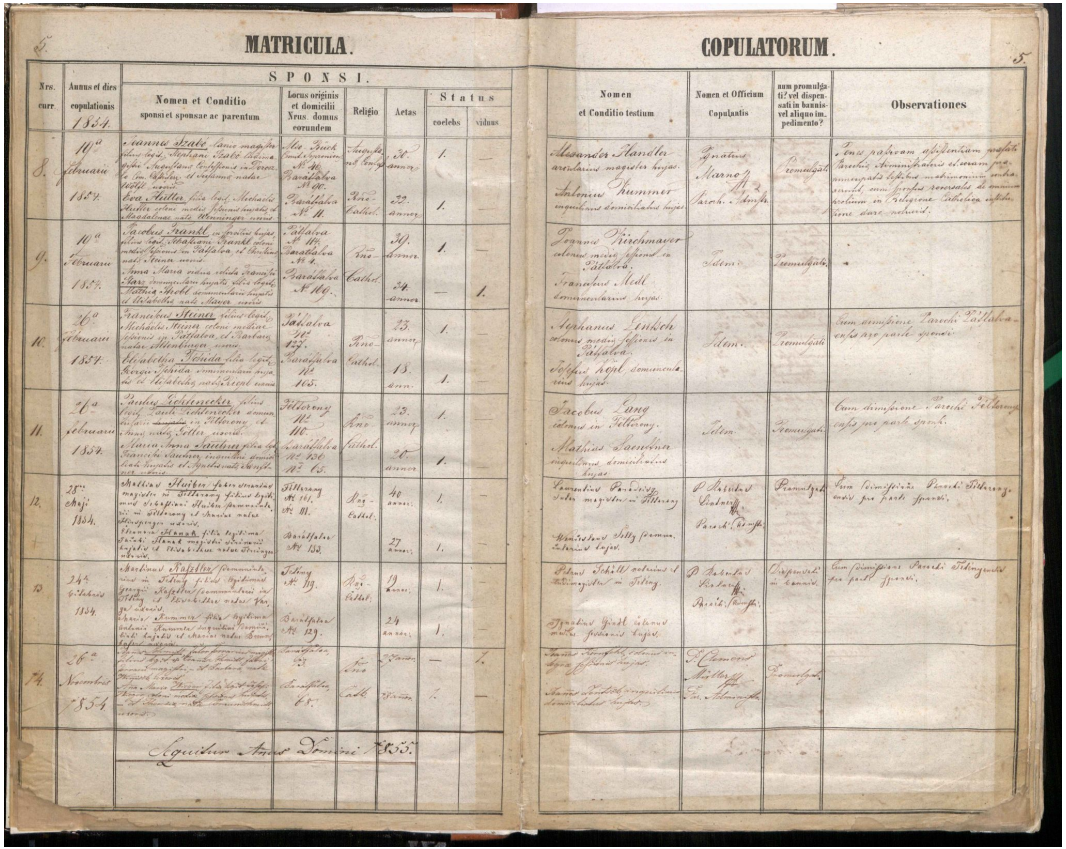
Transkribus (source: https://readcoop.eu/transkribus/)



Challenges

Images

- Tables are not uniform in structure (variation in fields/columns), owing to the range of periods and geographic locations the dataset comes from
- Handwriting styles vary due to differences in human-writers and periods
- Noise present in the form of ink blots, parchment texture, or general wear and tear of the physical documents

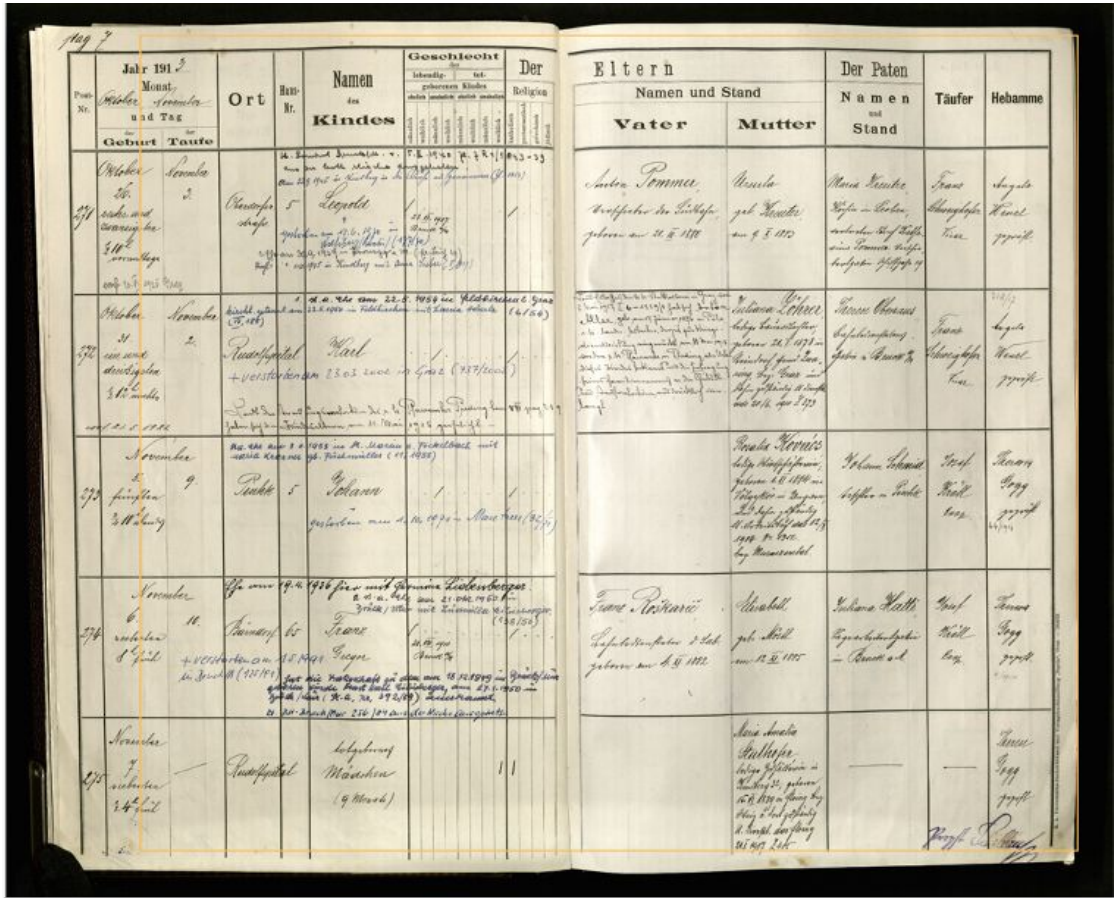


A parish register from Austria, 1855 (source: https://data.matricula-online.eu/)

Challenges

Images

- Poor contrast
- Overlapping text across columns
- Disregard any miscellaneous scribbles/notes



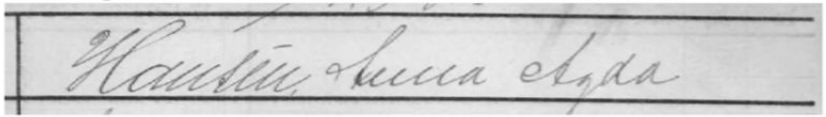
A parish birth register from Austria (source: https://data.matricula-online.eu/)

Using custom-made deep learning models for OCR

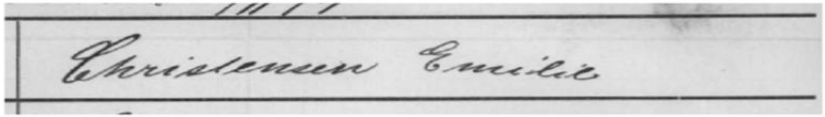
- Fine-tuning deep neural networks requires large amounts of data (>1 million images with labels)

- HAndwritten NAme (HANA) database:
  - Danish Census data, 1890-1923
  - Handwriting looks similar to that in our parish records
  - **~1.1 million** handwritten names
  - Accurate labels and clean images

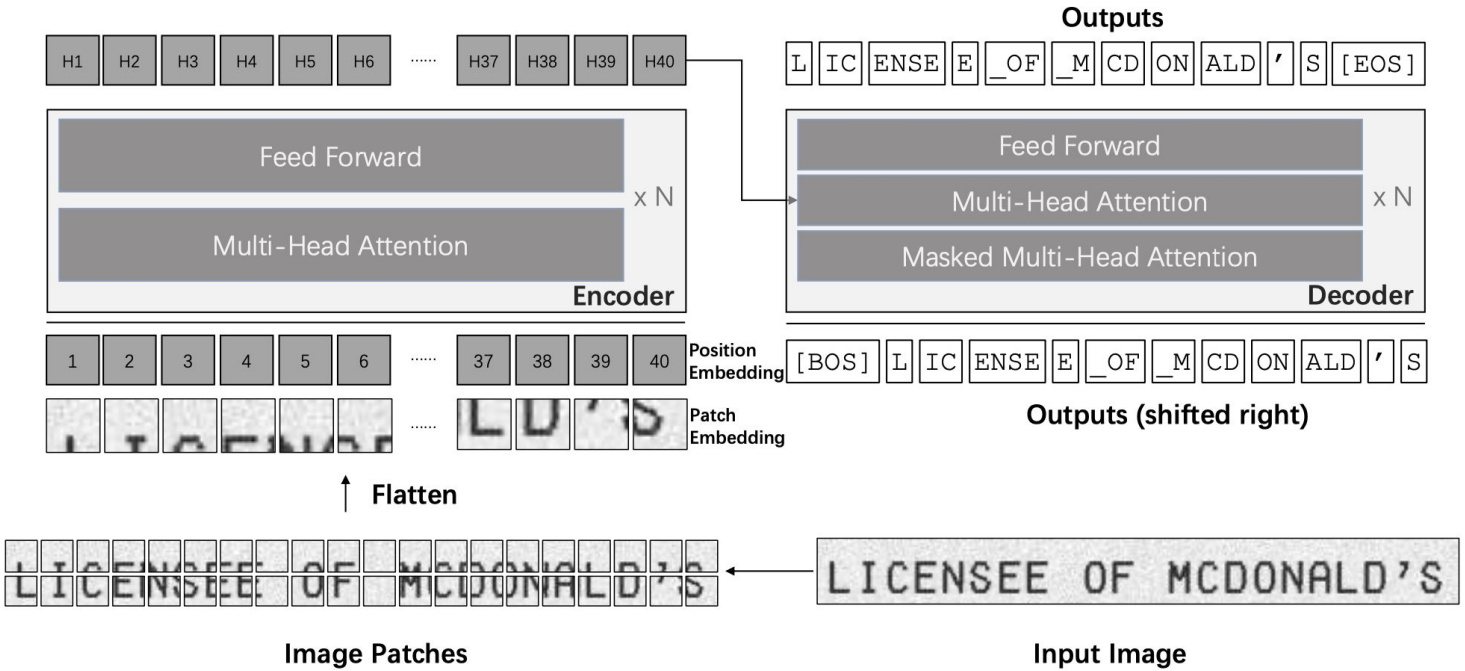
anna agda hansen



emilie christensen



TrOCR: Transformer-based OCR model



Resources utilised

- Monash M3 MASSIVE high-performance computing (HPC) cluster
  - Provides GPU nodes for running DL models
  - Large temporary storage and memory

- Amazon Web Services (AWS) (Commercial Cloud)
  - Lambda functions to run web-scrapers to fetch images from Matricula website
  - S3 buckets to store image repositories scraped from the web

Thank you!